Nonparametric learning and Regularization

Abstract

Several nonparametric methods in a regression model are presented. First, the most classical ones: piecewise polynomial estimators, estimation with Spline bases, kernel estimators and projection estimators on orthonormal bases (such as Fourier or wavelet bases). Since these methods suffer from the curse of dimensionality, we also present Generalized Additive Models and CART regression models.

The main references for this course are the following books :

- The elements of Statistical Learning by T. Hastie et al [2].
- Introduction to nonparametric statistics (2009) by A. Tsybakov [4]
- Introduction to High-Dimensional Statistics by C. Giraud [1]
- Concentration inequalities and model selection by P. Massart [3]

1 Introduction

We consider the regression model :

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i, \quad i = 1, \dots n.$$

We assume that the variables $\boldsymbol{X}_i$ are in $\mathbb{R}^d$, and $Y_i \in \mathbb{R}$. We assume that $\boldsymbol{X}_i$ is deterministic, and that the variables ε_i are i.i.d., centered, with variance σ^2.

Without any assumption on the function f, we are in a nonparametric framework. We propose several methods to estimate the function f. We first consider piecewise polynomial estimators.

2 Piecewise polynomial estimators.

We assume in this chapter that the variables X_i belong to some compact set of $\mathbb{R}$, that we assume to be $[0, 1]$.

2.1 Constant piecewise estimators

We estimate the function f by a piecewise constant function on a partition of $[0, 1]$. (These estimators are analogous to histograms for density estimation and are called regressograms).
We divide $[0, 1]$ into D regular intervals with the same size :

$$I_{k,D} = \mathbf{1}_{]k/D,(k+1)/D]}, \ k = 0, \dots, D-1.$$

It is quite natural to estimate the function f on the interval $I_{k,D}$ by the mean of the values of Y_i such that $X_i \in I_{k,D}$, hence for all $x \in I_{k,D}$, we define

$$\hat{f}_D(x) = \frac{\sum_{i, X_i \in I_{k,D}} Y_i}{\sharp\{i, X_i \in I_{k,D}\}}$$

if $\sharp\{i, X_i \in I_{k,D}\} \neq 0$ et

$$\hat{f}_D(x) = 0 \text{ if } \sharp\{i, X_i \in I_{k,D}\} = 0.$$

We can also write $\hat{f}_D(x)$ as follows

$$\hat{f}_D(x) = \frac{\sum_{i=1}^n Y_i \mathbf{1}_{X_i \in I_{k,D}}}{\sum_{i=1}^n \mathbf{1}_{X_i \in I_{k,D}}}.$$

In the following, we assume that $D < n$. If for all i, $X_i = i/n$, we have for all k, $\sharp\{i, X_i \in I_{k,D}\} \neq 0$.

This estimator corresponds to the least square estimator of f on the parametric model of constant piecewise functions on the intervals $I_{k,D}$:

$$\mathcal{S}_D = \{f(x) = \sum_{k=1}^{D} a_k \mathbf{1}_{x \in I_{k,D}}\}.$$

Indeed, if we minimize

$$h(a_1,\dots,a_D)=\sum_{i=1}^{n}\left(Y_i-\sum_{k=1}^{D}a_k\mathbf{1}_{X_i\in I_{k,D}}\right)^2=\sum_{k=1}^{D}\sum_{i,X_i\in I_{k,D}}(Y_i-a_k)^2, \tag{1}$$

the solution is

$$\hat{a}_l=\frac{\sum_{i,X_i\in I_{l,D}}Y_i}{\sharp\{i,X_i\in I_{l,D}\}},\quad \forall l.$$

Exercise. — Prove this result.

2.2 Piecewise polynomials

Piecewise polynomials with degree m on the partition defined by the intervals $I_{k,D}, 1\leq k\leq D$ correspond to the minimization of the criterion :

$$\sum_{i=1}^{n}\left(Y_i-\sum_{k=1}^{D}(a_{k,0}+a_{k,1}X_i+\ldots+a_{k,m}X_i^m)\mathbf{1}_{X_i\in I_{k,D}}\right)^2$$
$$=\sum_{k=1}^{D}\sum_{i,X_i\in I_{k,D}}(Y_i-a_{k,0}-a_{k,1}X_i-\ldots-a_{k,m}X_i^m)^2.$$

On each interval $I_{k,D}$, we adjust a polynomial with degree m, by minimizing the least square criterion

$$\sum_{i,X_i\in I_{k,D}}(Y_i-a_{k,0}-a_{k,1}X_i-\ldots-a_{k,m}X_i^m)^2.$$

This corresponds to a linear model with respect to the parameters $(a_{k,0},\dots,a_{k,m})$, we therefore have an explicit solution.

Exercise. — Prove this result.

The parameters D and m have to be calibrated.

2.3 Calibration of the parameters

We consider the constant piecewise estimator and we want to calibrate the parameter D. Let us first consider two extreme cases :

- If D is of the same order as the number of observations n, we have a single point X_i in each interval $I_{k,D}$ and we estimate f by Y_i on each interval $I_{k,D}$. The estimator is very irregular, simply reproducing the data. We have overfitting : small bias but high variance.

- On the contrary, if $D=1$, we estimate f on $[0,1]$ by the mean of all the observations Y_i. If f is far to be a constant function, the estimator is poor, it is underfitted, we have a large bias, and a small variance.

We have to find a compromise between these two extreme situations to realize a good bias/variance trade-off.

2.4 Theoretical performances of the estimator.

We assume that the regression function f is a Lipschitz function, this means that it belongs to the class of functions

$$\mathcal{S}_{1,R}=\{f\in\mathbb{L}^2([0,1]),\forall x,y\in[0,1],|f(x)-f(y)|\leq R|x-y|\}.$$

In this case, it is possible to give an upper bound for the quadratic risk of the estimator, for a suitable choice D.

THEOREM 1. — *We consider the regression model*

$$Y_i=f(\frac{i}{n})+\varepsilon_i,\ \ i=1,\dots,n.$$

The estimator

$$\hat{f}_D(x)=\frac{\sum_{i=1}^{n}Y_i\mathbf{1}_{X_i\in I_{k,D}}}{\sum_{i=1}^{n}\mathbf{1}_{X_i\in I_{k,D}}},$$

with

$$D=D(n)=[(nR^2)^{1/3}]$$

satisfies

$$\sup_{f\in\mathcal{S}_{1,R}}\mathbb{E}_f[\|\hat{f}_D-f\|_2^2]\leq C(\sigma)R^{\frac{2}{3}}n^{-\frac{2}{3}}$$

and the rate of convergence that is obtained is optimal.

Of course, this is a theoretical result, since in practice, it is impossible to know if the regression function f belongs to the class $\mathcal{S}_{1,R}$. We will see in Section 9 practical methods based on cross-validation to choose D.

Proof. —

- **Computation of the expectation**

  For all $x \in I_{k,D}$,

$$\mathbb{E}_f\left(\hat{f}_D(x)\right) = \frac{\sum_{i,X_i \in I_{k,D}} f(X_i)}{\sharp\{i, X_i \in I_{k,D}\}}.$$

$$\mathbb{E}_f\left(\hat{f}_D(x)\right) - f(x) = \frac{\sum_{i,X_i \in I_{k,D}} (f(X_i) - f(x))}{\sharp\{i, X_i \in I_{k,D}\}}.$$

  Assuming that $f \in \mathcal{S}_{1,R}$, we have for all x and X_i in the same interval $I_{k,D}$, $|x - X_i| \leq 1/D$, which implies that $|f(x) - f(X_i)| \leq RD^{-1}$. Hence,

$$|\text{Bias}(\hat{f}_D(x))| = |\mathbb{E}_f\left(\hat{f}_D(x)\right) - f(x)| \leq \frac{R}{D}.$$

- **Computation of the variance**

$$\begin{aligned}\text{Var}(\hat{f}_D(x)) &= \mathbb{E}_f[(\hat{f}_D(x) - \mathbb{E}_f(\hat{f}_D(x)))^2] \\ &= \frac{\sigma^2}{\sharp\{i, X_i \in I_{k,D}\}}.\end{aligned}$$

We consider the $\mathbb{L}^2([0,1], dx)$ risk to compute the performances of our estimator

$$L(\hat{f}_D, f) = \mathbb{E}_f[\int_0^1 (\hat{f}_D(x) - f(x))^2 dx].$$

We also have

$$L(\hat{f}_D, f) = \int_0^1 \mathbb{E}_f[(\hat{f}_D(x) - f(x))^2] dx.$$

Moreover,

$$\begin{aligned}\mathbb{E}_f[(\hat{f}_D(x) - f(x))^2] &= \mathbb{E}_f[\Big(\hat{f}_D(x) - \mathbb{E}_f(\hat{f}_D(x)) + \mathbb{E}_f(\hat{f}_D(x)) - f(x)\Big)^2] \\ &= \mathbb{E}_f[(\hat{f}_D(x) - \mathbb{E}_f(\hat{f}_D(x)))^2] + [\mathbb{E}_f(\hat{f}_D(x)) - f(x)]^2 \\ &= \text{Var}(\hat{f}_D(x)) + \text{Bias}^2(\hat{f}_D(x)) \\ &\leq \frac{\sigma^2}{\sharp\{i, X_i \in I_{k,D}\}} + R^2 D^{-2}.\end{aligned}$$

Since $X_i = i/n$, we remark easily that $\sharp\{i, X_i \in I_{k,D}\} \geq [n/D] \geq n/(2D)$ assuming that $D \leq n/2$. This implies:

$$L(\hat{f}_D, f) \leq \frac{2\sigma^2 D}{n} + R^2 D^{-2}.$$

It remains to choose D to optimize this quadratic risk. We set

$$D = [(nR^2)^{1/3}],$$

and we obtain

$$L(\hat{f}_D, f) \leq C(\sigma) R^{\frac{2}{3}} n^{-\frac{2}{3}}.$$

The proof of the lower bounds are given in Tsybakov[4]. ∎

3 Splines

We assume that $X_i \in \mathbb{R}$. The estimators proposed in the previous section are not continuous; in order to get estimators that are piecewise polynomial, with regularity properties, we use Spline bases.

3.1 Linear and cubic Splines

We will consider estimators that are piecewise linear functions of the form

$$f(x) = \beta_0 + \beta_1 x + \beta_2 (x-a)_+ + \beta_3 (x-b)_+ + \beta_4 (x-c)_+ + \ldots +$$

where $0 < a < b < c \ldots$ are the points that define the partition (also called nodes), and $x_+ = \max(x, 0)$.

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x \text{ if } x \leq a \\ &= \beta_0 + \beta_1 x + \beta_2 (x-a)_+ \text{ if } a \leq x \leq b \\ &= \beta_0 + \beta_1 x + \beta_2 (x-a)_+ + \beta_3 (x-b)_+ \text{ if } b \leq x \leq c \end{aligned}$$

The function f is continuous, if we want a more regular estimator (for example a twice continuously differentiable estimator), we consider cubic splines.

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-a)_+^3 + \beta_5 (x-b)_+^3 + \beta_4 (x-c)_+^3 + \ldots +$$

The function $(x-a)^3$ vanishes as well as its first and second derivatives in a, hence f is twice continuously differentiable.

In order to avoid problems on the boundaries, we generally impose additional constraints on cubic splines, namely that the function is linear on the two boundary intervals corresponding to the ends.
Assume that we are on $[0,1]$. $\xi_0 = 0 < \xi_1 < \ldots < \xi_K < 1$.

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3.$$

We require that $f''(0) = f^{(3)}(0) = 0$, $f''(\xi_K) = f^{(3)}(\xi_K) = 0$. Hence we have:

$$\beta_2 = \beta_3 = 0, \ \sum_{k=1}^{K} \theta_k (\xi_K - \xi_k) = 0, \ \sum_{k=1}^{K} \theta_k = 0.$$

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \sum_{k=1}^{K} \theta_k [(x-\xi_k)_+^3 - (x-\xi_K)_+^3] \\ &= \beta_0 + \beta_1 x + \sum_{k=1}^{K-1} \theta_k (\xi_K - \xi_k) [\frac{(x-\xi_k)_+^3 - (x-\xi_K)_+^3}{(\xi_K - \xi_k)}] \end{aligned}$$

We set $\gamma_k = \theta_k (\xi_K - \xi_k)$ and $d_k(x) = \frac{(x-\xi_k)_+^3 - (x-\xi_K)_+^3}{(\xi_K - \xi_k)}$. $\sum_{k=1}^{K-1} \gamma_k = 0$.

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \gamma_k (d_k(x) - d_{K-1}(x)).$$

We obtain the *natural cubic splines* basis :

$$N_1(x) = 1, N_2(x) = x, \ \forall 1 \leq k \leq K-2, N_{k+2}(x) = d_k(x) - d_{K-1}(x).$$

We have to choose the position and the number of nodes.

3.2 Regularization methods for cubic splines

We consider the regression model : $Y_i = f(X_i) + \epsilon_i, \ 1 \leq i \leq n$.
We minimize among the functions f that are natural cubic splines with nodes at the observation points X_i ($f(x) = \sum_{k=1}^{n} \theta_k N_k(x)$) the penalized criterion:

$$C(f, \lambda) = \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \int_0^1 (f''(t))^2 dt,$$

where $\lambda > 0$. If we denote $\Omega_{l,k} = \int_0^1 N_k''(x) N_l''(x) dx$ and $N_{i,j} = N_j(X_i)$, the criterion to minimize is

$$C(\theta, \lambda) = \|Y - N\theta\|^2 + \lambda \theta^* \Omega \theta.$$

The solution is

$$\hat{\theta} = (N^* N + \lambda \Omega)^{-1} N^* Y$$

where

$$\hat{f}(x) = \sum_{k=1}^{n} \hat{\theta}_k N_k(x). \tag{2}$$

Exercise. — Prove this result.

THEOREM 2. — *We denote by*

$$\mathcal{F} = \{f, \ C^2([0,1]), \int_0^1 f''^2(t) dt < +\infty\}.$$

Let $n \geq 2$, $0 < X_1 < \ldots < X_n < 1$ *and* $(Y_1, \ldots, Y_n) \in \mathbb{R}^n$. *For* $f \in \mathcal{F}$, *and* $\lambda > 0$, *we denote by*

$$C(f, \lambda) = \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \int_0^1 (f''(t))^2 dt.$$

For any $\lambda > 0$, there exists a unique minimizer f in $\mathcal{F}$ of the criterion $C(f, \lambda)$, which is the function $\hat{f}$ defined in (2).

This theorem is very powerful and shows that the function $\hat{f}$ is a minimizer of the criterion $f \mapsto C(f, \lambda)$ over a much larger class of functions than the cubic splines with nodes at the observation points X_i since it is indeed a minimizer of the criterion over the class $\mathcal{F}$.

4 Kernel estimators

We consider the regression model

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{3}$$

where $\boldsymbol{X}_i \in \mathbb{R}^d$, and the variables ε_i are i.i.d. centered, with variance σ^2, The variables $\boldsymbol{X}_i$ may be random, in this case, they are independent of the variables ε_i.

4.1 Definition of the kernel estimator

DEFINITION 3. — *We call Kernel a function $K : \mathbb{R}^d \to \mathbb{R}$ such that $\int K^2 < +\infty$ and $\int K = 1$.*

DEFINITION 4. — *We introduce a positive parameter $h > 0$ (that we call window) and a kernel K. The kernel estimator of f in Model (3) associated to the kernel K and the window h is the function $\hat{f}_h$ defined by :*

$$\hat{f}_h(\boldsymbol{x}) = \frac{\sum_{i=1}^n Y_i K\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right)}.$$

When the $\boldsymbol{X}_i$'s are i.i.d. with uniform distribution on $[0,1]^d$, we also find the following definition :

$$\hat{f}_h(\boldsymbol{x}) = \frac{1}{nh^d} \sum_{i=1}^n Y_i K\left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right). \tag{4}$$

If, for example $d = 1$ and $K(u) = (1/2)\mathbf{1}_{|u|\leq 1}$, $\hat{f}_h(\boldsymbol{x})$ is the mean of the values Y_i such that $|\boldsymbol{X}_i - \boldsymbol{x}| \leq h$. This is a constant piecewise estimator.

Extreme cases :
Assume that $d = 1$ and that the X_i's are regularly spaced on $[0, 1]$.
-If $h = 1/n$, the estimator is very irregular and reproduces the data.
-If $h \geq 1$, for all x, $\hat{f}_h(x) = \sum_{i=1}^n Y_i/n$.

Here again, we have to optimize the value of the window h to realize a good compromise between the bias term and the variance term.

Remark : we generally use regular kernels, leading to regular estimators.
Examples of kernels in dimension 1 :
-The triangular kernel $K(x) = (1 - |x|)\mathbf{1}_{|x|\leq 1}$.
-The Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
-The parabolic kernel $K(x) = \frac{3}{4}(1 - x^2)\mathbf{1}_{|x|\leq 1}$.

4.2 Theoretical properties of the kernel estimators.

For the sake of simplicity, we consider the model where the X_i's are random, i.i.d. with uniform distribution on $[0, 1]$ and we consider the estimator defined by (4).

THEOREM 5. — *Assume that the regression function f belongs to the class $\Sigma(\beta, R)$ defined by*

$$\Sigma(\beta, R) = \left\{ f \in \mathcal{C}^l([0,1]), \forall x, y \in [0,1], |f^{(l)}(x) - f^{(l)}(y)| \leq R|x - y|^\alpha \right\},$$

where $\beta = l + \alpha$ with $l \in \mathbb{N}$ and $\alpha \in]0, 1]$.
We make the following assumptions on the kernel K :
H1 $\int u^j K(u)du = 0$ *for* $j = 1, \ldots, l$.
H2 $\int |u|^\beta |K(u)|du < +\infty$.
If we choose h such that $h \approx (nR^2)^{-1/(1+2\beta)}$, we get, $\forall f \in \Sigma(\beta, R)$,

$$\mathbb{E}_f \left(\int_0^1 (\hat{f}_h(x) - f(x))^2 \right) \leq C(\beta, \sigma, \|s\|_\infty) R^{\frac{2}{1+2\beta}} n^{-\frac{2\beta}{1+2\beta}},$$

and this rate of convergence is optimal.

Proof. —

Computation of the bias : we denote $K_h = (1/h)K(./h)$,

$$\mathbb{E}_f(\hat{f}_h(x)) = \int_0^1 f(y)K_h(x-y)dy = f \star K_h(x).$$

Hence, since $\int K = 1$, we get

$$\mathbb{E}_f(\hat{f}_h(x)) - f(x) = \int (f(x-uh) - f(x))K(u)du.$$

We use a Taylor expansion :

$$f(x-uh) = f(x) - f'(x)uh + f''(x)\frac{(uh)^2}{2} + \ldots + f^{(l)}(x-\tau uh)\frac{(-uh)^l}{l!}$$

with $0 \leq \tau \leq 1$. Using Assumption **H1**,

$$\begin{aligned}\mathbb{E}_f(\hat{f}_h(x)) - f(x) &= \int f^{(l)}(x-\tau uh)\frac{(-uh)^l}{l!}K(u)du \\ &= \int (f^{(l)}(x-\tau uh) - f^{(l)}(x))\frac{(-uh)^l}{l!}K(u)du.\end{aligned}$$

Since $f \in \Sigma(\beta, R)$, and using Assumption **H2**, we get

$$|\mathbb{E}_f(\hat{f}_h(x)) - f(x)| \leq R\tau^\alpha h^\beta \frac{1}{l!}\int |u|^\beta |K(u)|du.$$

Computation of the variance :

$$\begin{aligned}Var(\hat{f}_h(x)) &= \frac{1}{n^2}\sum_{i=1}^n Var(Y_i K_h(x-X_i)). \\ &\leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_s[Y_i^2 K_h^2(x-X_i)] \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[f^2(X_i)K_h^2(x-X_i) + \varepsilon_i^2 K_h^2(x-X_i)].\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{E}[f^2(X_i)K_h^2(x-X_i)] &= \int f^2(y)\frac{1}{h^2}K^2(\frac{x-y}{h})dy \\ &= \int f^2(x-uh)\frac{1}{h}K^2(u)du \\ &\leq \|f\|_\infty^2 \frac{1}{h}\int K^2.\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\varepsilon_i^2 K_h^2(x-X_i)] &= \sigma^2 \int \frac{1}{h^2}K^2(\frac{x-y}{h})dy \\ &= \frac{\sigma^2}{h}\int K^2.\end{aligned}$$

Hence, we have

$$Var(\hat{f}_h(x)) \leq C(\|f\|_\infty, \sigma)\frac{1}{nh}.$$

Since

$$\mathbb{E}_f\left(\int_0^1 (\hat{f}_h(x) - f(x))^2 dx\right) = \int_0^1 \left(Biais^2(\hat{f}_h(x)) + Var(\hat{f}_h(x))\right)dx,$$

we get

$$\mathbb{E}_f\left(\int_0^1 (\hat{f}_h(x) - f(x))^2 dx\right) \leq C(\beta, \sigma, \|f\|_\infty)\left(R^2h^{2\beta} + \frac{1}{nh}\right).$$

If we choose h such that

$$R^2h^{2\beta} \approx \frac{1}{nh},$$

that is $h \approx (nR^2)^{-1/(1+2\beta)}$, we obtain the desired results. For the proof of the lower bounds, see Tsybakov[4]. ∎

5 Pointwise estimation by local polynomials

In Section 2, we have fixed a partition, this partition does not depend on the data. The estimation of the regression function f at point x was obtained from

the observations at the points X_i that lye in the same interval as the point x, which leads to irregular estimators. A natural idea is to estimate the function f at point x with the observations such that X_i is "close" to x. More generally, we introduce a weight function $(w_i(x))$ which derives from a kernel : $w_i(x) = K((X_i - x)/h)$ and that will place a greater weight in the observations for which X_i is "close" to x, and we minimize (with respect to a) the weighted sum of squares :

$$\sum_{i=1}^{n} w_i(x)(Y_i - a)^2.$$

The solution is given by

$$a = \hat{f}_n(x) = \frac{\sum_{i=1}^n w_i(x) Y_i}{\sum_{i=1}^n w_i(x)}, \tag{5}$$

which corresponds to the kernel estimator defined in previous section ! Hence, we simply have here a new interpretation of the kernel estimator. We can generalize the previous formula by replacing the constant a by a polynomial with degree p. Given a point x where we want to estimate the regression function, for u in a neighborhood of x, we consider the polynomial

$$P_x(u, a) = a_0 + a_1(u - x) + \ldots + \frac{a_p}{p!}(u - x)^p.$$

We want to estimate the regression function in a neighborhood of x by the polynomial $P_x(u, a)$ where the vector $a = (a_0, \ldots a_p)$ is obtained by minimizing the weighted sum of squares :

$$\sum_{i=1}^{n} w_i(x)(Y_i - a_0 - a_1(X_i - x) - \ldots - \frac{a_p}{p!}(X_i - x)^p)^2.$$

The solution is given by the vector $\hat{a}(x) = (\hat{a}_0(x), \ldots \hat{a}_p(x))$, the local estimator of the regression function f is

$$\hat{f}_n(u) = \hat{a}_0(x) + \hat{a}_1(x)(u - x) + \ldots + \frac{\hat{a}_p(x)}{p!}(u - x)^p.$$

At the point x, we get :

$$\hat{f}_n(x) = \hat{a}_0(x).$$

Note that this estimator does not correspond to the one obtained in (5), which is obtained for $p = 0$ (this is the kernel estimator). If $p = 1$, this method is called the *local linear regression*. We can express the value of $\hat{a}_0(x)$ from a weighted least square criterion : let X_x denote the matrix

$$X_x = \begin{pmatrix} 1 & X_1 - x & \ldots & \frac{(X_1-x)^p}{p!} \\ 1 & X_2 - x & \ldots & \frac{(X_2-x)^p}{p!} \\ . & . & . & . \\ . & . & . & . \\ 1 & X_n - x & \ldots & \frac{(X_n-x)^p}{p!} \end{pmatrix}.$$

Let W_x the diagonal matrix with ith component on the diagonal $w_i(x)$. We therefore have:

$$\sum_{i=1}^{n} w_i(x)(Y_i - a_0 - a_1(X_i - x) - \ldots - \frac{a_p}{p!}(X_i - x)^p)^2 = (Y - X_x a)^* W_x (Y - X_x a).$$

Minimizing the above expression leads to the weighted least square estimator:

$$\hat{a}(x) = (X_x^* W_x X_x)^{-1} X_x^* W_x Y,$$

and the local polynomial estimator at point x corresponds to $\hat{f}_n(x) = \hat{a}_0(x)$, which is the scalar product of the vector Y with the first line of the matrix $(X_x^* W_x X_x)^{-1} X_x^* W_x$. We have the following theorem :

THEOREM 6. — *The local polynomial estimator at point x is*

$$\hat{f}_n(x) = \sum_{i=1}^{n} l_i(x) Y_i$$

where $l(x)^* = (l_1(x), \ldots, l_n(x))$,

$$l(x)^* = e_1^* (X_x^* W_x X_x)^{-1} X_x^* W_x,$$

with $e_1^* = (1, 0, \ldots, 0)$.

$$\mathbb{E}(\hat{f}_n(x)) = \sum_{i=1}^{n} l_i(x) f(X_i)$$

$$Var(\hat{f}_n(x)) = \sigma^2 \sum_{i=1}^{n} l_i^2(x).$$

6 Projection estimators

We consider the regression model

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n. \tag{6}$$

Let $(\phi_j, j \geq 1)$ an orthonormal basis of $\mathbb{L}^2([0,1])$. For $D \geq 1$ we define

$$S_D = \text{Vect}\{\phi_1, \dots, \phi_D\}.$$

We denote by f_D the orthogonal projection of f onto S_D in $\mathbb{L}^2([0,1])$:

$$f_D = \sum_{j=1}^{D} \langle f, \phi_j \rangle \phi_j,$$

where

$$\theta_j = \langle f, \phi_j \rangle = \int_0^1 f(x)\phi_j(x)dx.$$

We estimate θ_j by

$$\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_j(X_i).$$

Indeed, if the X_i's are deterministic,

$$\mathbb{E}(\hat{\theta}_j) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)\phi_j(X_i),$$

and if $f\phi_j$ is regular and the X_i's are equispaced on $[0,1]$, this quantity is close to θ_j. If the X_i's are random, with uniform distribution on $[0,1]$, we have

$$\mathbb{E}(\hat{\theta}_j) = \theta_j.$$

We introduce the estimator

$$\hat{f}_D(x) = \sum_{j=1}^{D} \hat{\theta}_j \phi_j(x),$$

which is called *estimator by projection.*

Example of the Fourier basis

We denote by $(\phi_j, j \geq 1)$ the Fourier basis of $\mathbb{L}^2([0,1])$:

$$\begin{aligned}
&\phi_1(x) = \mathbf{1}_{[0,1]},\\
&\phi_{2k}(x) = \sqrt{2}\cos(2\pi kx) \quad \forall k \geq 1\\
&\phi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx) \quad \forall k \geq 1.
\end{aligned}$$

We obtain for all $D \geq 1$, the estimator

$$\hat{f}_D(x) = \frac{1}{n}\sum_{j=1}^{D}\sum_{i=1}^{n} Y_i \phi_j(X_i)\phi_j(x).$$

We can prove theoretical performances of the estimator, assuming that the regression function f belongs to a class of regular and periodical functions.

DEFINITION 7. — *Let $L > 0$ and $\beta = l + \alpha$ with $l \in \mathbb{N}$ and $\alpha \in]0,1]$. We define the class $\Sigma^{per}(\beta, R)$ by*

$$\Sigma^{per}(\beta, R) = \Big\{ f \in \mathcal{C}^l([0,1]), \forall j = 0, \dots, l, \quad f^{(j)}(0) = f^{(j)}(1),$$

$$\forall x, y \in [0,1], |f^{(l)}(x) - f^{(l)}(y)| \leq R|x-y|^{\alpha} \Big\}.$$

THEOREM 8. — *In the model*

$$Y_i = f(\frac{i}{n}) + \varepsilon_i, \quad i = 1, \dots, n,$$

where the ε_i's are i.i.d. with distribution $\mathcal{N}(0, \sigma^2)$, the estimator $\hat{f}_D$ defined for all $x \in [0,1]$ by :

$$\hat{f}_D(x) = \frac{1}{n}\sum_{j=1}^{D}\sum_{i=1}^{n} Y_i \phi_j(X_i)\phi_j(x)$$

with $D = [(nR^2)^{1/(1+2\beta)}]$, satisfies far all $\beta > 1$, $R > 0$,

$$\sup_{f \in \Sigma^{per}(\beta,R)} \mathbb{E}_f\left(\|\hat{f}_D - f\|_2^2\right) \leq C(\beta, \sigma) R^{\frac{2}{1+2\beta}} n^{\frac{-2\beta}{1+2\beta}}.$$

In the next chapter, we introduce wavelet bases, that are used in particular to estimate very irregular functions.

7 Wavelet bases and regularization

In this chapter, we consider the problem of estimating spatially inhomogeneous functions, namely, functions that may be very regular in some parts of their definition domain, and very irregular (presenting peaks) in other parts of the space. Wavelet bases are orthonormal bases, which are well adapted to estimate this type of functions. We assume here that the X_i's are in $[0,1]$. In the practical works, we will consider one dimensional examples for signal processing and we will also consider two-dimensional examples for image processing.

7.1 Wavelet bases

Haar basis

The Haar basis is the most simple wavelet basis. The father wavelet (or scaling function) is defined by

$$\begin{aligned}\phi(x) &= 1 \text{ si } x \in [0,1[,\\ &= 0 \text{ sinon.}\end{aligned}$$

The mother wavelet (or wavelet function) is defined by

$$\begin{aligned}\psi(x) &= -1 \text{ si } x \in [0,1/2],\\ &= 1 \text{ si } x \in]1/2,1].\end{aligned}$$

For all $j \in \mathbb{N}$, $k \in \mathbb{N}$, we define

$$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k), \quad \psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k).$$

THEOREM 9. — *The functions* $(\phi, \psi_{j,k}, j \in \mathbb{N}, k \in \{0,\dots,2^j-1\})$ *form an orthonormal basis of* $\mathbb{L}^2([0,1])$.

We deduce from this theorem that one can expand a function belonging to $\mathbb{L}^2([0,1])$ in this basis :

$$f(x) = \alpha\phi(x) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k}(x).$$

$\alpha = \int_0^1 f(x)\phi(x)dx$ is called "scaling coefficients" and the $\beta_{j,k} = \int_0^1 f(x)\psi_{j,k}(x)dx$ are called "details". The approximation of f at the resolution level J is the function

$$f_J = \alpha\phi(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k}(x).$$

This expression includes 2^J coefficients. As the space generated by the functions $(\phi, \psi_{j,k}, 0 \le j \le J-1, 0 \le k \le 2^j-1)$ is the space of constant piecewise functions over the intervals of length $1/2^J$, that is the space generated by the functions $(\phi_{J,k}, 0 \le k \le 2^J-1)$, we also have

$$f_J = \sum_{k=0}^{2^J-1}\alpha_{J,k}\phi_{J,k}(x),$$

where $\alpha_{J,k} = \int_0^1 f(x)\phi_{J,k}(x)dx$.

The Haar basis is easy to define, its functions are compactly supported. Nevertheless, this basis leads to irregular approximations : the projection are not even continuous. There exist other wavelet bases, that are compactly supported and more regular, for example Daubechies wavelets. (See Daubechies (1992) : *Ten Lectures on wavelets*).

7.2 Estimating a regression function with wavelet's projections

Wavelets are well adapted for the analysis of signals, sampled on a regular, dyadic grid. They are widely used for signal and image processing. We consider the model :

$$Y_k = f(\frac{k}{N}) + \epsilon_k,\ k = 1,\dots,N = 2^J.$$

We consider the $N = 2^J$ first functions of a wavelet basis on $[0,1]$: $(\phi, \psi_{j,k}, 0 \le j \le J-1, 0 \le k \le 2^j-1)$. We denote by W the $N * N$

matrix

$$W = \frac{1}{\sqrt{N}} \begin{pmatrix} \phi(1/N) & \psi_{0,0}(1/N) & \dots & \psi_{J-1,2^J-1}(1/N) \\ . & . & . & . \\ \phi(i/N) & \psi_{0,0}(i/N) & \dots & \psi_{J-1,2^J-1}(i/N) \\ . & . & . & . \\ \phi(N/N) & \psi_{0,0}(N/N) & \dots & \psi_{J-1,2^J-1}(N/N) \end{pmatrix}$$

In the case of the Haar basis, W is an orthogonal matrix (the Haar basis is also orthonormal for the discrete scalar product). We denote by W^* the transpose of W and

$$\hat{\theta} = W^* Y,$$

the wavelet transform of the vector Y.
This is also the least square estimator of θ in the model $Y = W\theta + \varepsilon$ since W is orthogonal.

$$\begin{aligned} \hat{\theta}_{j,k} = \frac{1}{\sqrt{N}} \sum_{l=1}^{N} \psi_{j,k}(\frac{l}{N}) Y_l &= \frac{1}{\sqrt{N}} \sum_{l=1}^{N} \psi_{j,k}(\frac{l}{N}) f(\frac{l}{N}) + \tilde{\epsilon}_l \\ &\approx \sqrt{N}\beta_{j,k} + \tilde{\epsilon}_l \end{aligned}$$

where

$$\begin{aligned} \tilde{\epsilon}_l &= \frac{1}{\sqrt{N}} \sum_{l=1}^{N} \psi_{j,k}(\frac{l}{N}) \epsilon_l \\ &\sim \mathcal{N}(0, \frac{\sigma^2}{N} \sum_{l=1}^{N} \psi_{j,k}^2(\frac{l}{N})). \end{aligned}$$

In the case of the Haar basis, $\frac{\sigma^2}{N} \sum_{l=1}^{N} \psi_{j,k}^2(\frac{l}{N}) = \sigma^2$. We can recover a signal from its wavelet transform by the inverse transform :

$$Y = (W^*)^{-1} \hat{\theta}.$$

$Y = W\hat{\theta}$ in the case of the Haar basis.

Denoising by linear approximation

We approximate the regression function f by the orthogonal projection of f onto V_{J_0} :

$$f_{J_0} = \alpha\phi + \sum_{j=0}^{J_0-1} \sum_{k=0}^{2^j-1} \beta_{j,k} \psi_{j,k},$$

which corresponds to keep only the 2^{J_0} first wavelet coefficients. To estimate f_{J_0}, we keep only the 2^{J_0} first coefficients in the vector $\hat{\theta}$, the other coefficients ar set to 0. This defines a vector that we denote $\hat{\theta}_{J_0}$, then we reconstruct the denoised signal

$$\hat{Y}_{J_0} = (W^*)^{-1} \hat{\theta}_{J_0}.$$

The regression function f is finally estimated by

$$\hat{f}_{J_0}(x) = \frac{1}{\sqrt{N}} (\phi(x), \psi_{0,0}(x), \dots, \psi_{J-1,2^J-1}(x)) \hat{\theta}_{J_0}.$$

$$\hat{f}_{J_0}(x) = \hat{\alpha}\phi(x) + \sum_{j=0}^{J_0-1} \sum_{k=0}^{2^j-1} \hat{\beta}_{j,k} \psi_{j,k}(x)$$

where $\hat{\theta}_{J_0} = \sqrt{N}(\hat{\alpha}, \hat{\beta}_{j,k}, j = 0, \dots J_0 - 1, k = 0, \dots, 2^j - 1, 0, \dots, 0)$. We have to choose the parameter J_0 in an optimal way.

Denoising by nonlinear approximation via thresholding

The thresholding method relies on the minimization with respect to $\theta \in \mathbb{R}^N$ of the penalized least square criterion with l_1 penalty

$$C(\theta) = \|Y - W\theta\|^2 + 2\lambda\|\theta\|_1,$$

with $\|\theta\|_1 = \sum_{i=1}^{N} |\theta_i|$. If W is orthogonal (we recall that this is the case for the Haar basis in particular), this leads to an explicit solution. The solution is

$$\begin{aligned} |\tilde{\theta}_i| &= |\hat{\theta}_i| - \lambda \text{ si } |\hat{\theta}_i| \geq \lambda \\ &= 0 \text{ si } |\hat{\theta}_i| \leq \lambda \end{aligned}$$

$$\tilde{\theta}_i = \text{sign}(\hat{\theta}_i)(|\hat{\theta}_i| - \lambda)\mathbf{1}_{|\hat{\theta}_i| \geq \lambda}.$$

Exercise. — Prove this result.

This method is called "soft thresholding", we apply a continuous function to $\hat{\theta}_i$. The "hard thresholding" consists in setting

$$\tilde{\theta}_i = \hat{\theta}_i \mathbf{1}_{|\hat{\theta}_i| \geq \lambda}.$$

we reconstruct the denoised signal

$$\tilde{Y} = W\tilde{\theta}.$$

The regression function f is estimated by

$$\hat{f}_N(x) = \frac{1}{\sqrt{N}}(\phi(x), \psi_{0,0}(x), \ldots, \psi_{J-1,2^J-1}(x))\tilde{\theta}.$$

We denote $\tilde{\theta} = \sqrt{N}(\tilde{\alpha}, \tilde{\beta}_{j,k}, j = 0, \ldots, J-1, k = 0, \ldots 2^{j-1})$, we obtain

$$\hat{f}_N(x) = \tilde{\alpha}\phi(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \tilde{\beta}_{j,k}\psi_{j,k}(x).$$

In practice, we have to choose the threshold λ. We take generally $\lambda = \sigma\sqrt{2\log(N)}$. Indeed,

$$\hat{\theta} = W^*Y = \frac{1}{\sqrt{N}}\sum_{l=1}^{N} \psi_{j,k}(\frac{l}{N})f(\frac{l}{N}) + \tilde{\epsilon}_l$$

with

$$\tilde{\epsilon} = W^*\epsilon \sim \mathcal{N}_N(0, \sigma^2 I_N).$$

One can prove that

$$E\left(\sup_{1\leq i\leq N} |\tilde{\epsilon}_i|\right) \approx \sigma\sqrt{2\log(N)}.$$

The coefficients that are smaller to $\sigma\sqrt{2\log(N)}$ are considered as noise, and set to 0. These thresholding methods allow to estimate very irregular signals (in particular, functions with peaks, that can be represented with a small number of wavelet coefficients after thresholding).

8 Generalized additive models

The previous methods suffer from the curse of dimensionality and are mostly used in dimension 1. They are based on local means of the observations, and we have seen, that, in large dimensions, observations are isolated. Under additional hypotheses on the structure of the regression function, one can get around this problem. We consider in this chapter additive regression functions. We introduce the additive model

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i,$$

where the ε_i's are i.i.d. centered, with variance σ^2, and $\boldsymbol{X}_i \in \mathbb{R}^d$. We assume that the regression function f is additive (or can be well approximated by an additive function), namely that

$$f(\boldsymbol{X}_{i,1}, \ldots, \boldsymbol{X}_{i,d}) = \alpha + f_1(\boldsymbol{X}_{i,1}) + \ldots + f_d(\boldsymbol{X}_{i,d}).$$

In order to guaranty the unicity of this expression, we require that

$$\int_{\mathbb{R}} f_j(x_j)dx_j = 0, \quad \forall j = 1, \ldots, d.$$

We will describe in this section a method to estimate the components of an additive model, we call these models GAM (Generalized Additive Models). We assume that each one dimensional function is estimated with Spline bases as explained in Section 3.2. We introduce the penalized criterion

$$\begin{aligned}\text{Crit}(\alpha, f_1, f_2, \ldots, f_p) &= \sum_{i=1}^{n}\left(Y_i - \alpha - \sum_{j=1}^{d} f_j(X_{i,j})\right)^2 \\ &+ \sum_{j=1}^{d} \lambda_j \int (f_j'')^2(x_j)dx_j,\end{aligned}$$

where for all j, $\lambda_j \geq 0$ is a regularization parameter. One can prove that the solution of this minimization problem is an additive model of cubic splines, each function $\hat{f}_j$ is a cubic spline with respect to the variable x_j, whose nodes correspond to the different values of the variables $X_{i,j}, i = 1, \ldots n$. In order

to guaranty the unicity of the minimizer, we impose the following constraints :

$$\forall j = 1, \ldots, d, \quad \sum_{i=1}^{n} f_j(X_{i,j}) = 0.$$

Under these conditions, we get $\hat{\alpha} = \sum_{i=1}^{n} Y_i/n$, and if the matrix with entries $X_{i,j}$ is not singular, one can show that the criterion is strictly convex, and admits therefore a unique minimizer. The following algorithm, called the backfitting algorithm, converges to the solution :

Backfitting algorithm for GAM models :

1. Initialization : $\hat{\alpha} = \sum_{i=1}^{n} Y_i/n$, $\hat{f}_j = 0 \; \forall j$.
2. For $l = 1$ to Niter
   For $j = 1$ to d
   - $\hat{f}_j$ minimize

   $$\sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(X_{i,k}) - f_j(X_{i,j}) \right)^2 + \lambda_j \int (f_j'')^2(x_j) dx_j,$$

   - $\hat{f}_j := \hat{f}_j - \frac{1}{n} \sum_{i=1}^{n} \hat{f}_j(X_{i,j})$.

   Stop when all the functions $\hat{f}_j$ are "stable".

The same algorithm can be used with other estimation methods than cubic splines such as local polynomials, kernel estimators, projection estimators .. The generalized additive models form an extension to linear models, they are more flexible, but still very easy to interpret. These models are widely used for statistical modeling. Nevertheless, in large dimension, they may be hard to implement and it may be useful to combine them with a selection algorithm, in order to reduce the dimension.

9 Regression trees CART

The CART algorithm (Classification And Regression Trees) is also a non parametric method to build estimators of a regression function in a multidimensional framework. The methods based on trees rely on a partition of the space of input variables, we then infer a simple model (for example constant piecewise functions) on each element of the partition. We assume that we have a n sample $(\boldsymbol{X}_i, Y_i)_{1 \leq i \leq n}$ with $\boldsymbol{X}_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. The CART algorithm allows to define, from the learning sample, an automatic data driven partition of the space of input variables $\boldsymbol{X}_i$. Assume that the space of input variables $\boldsymbol{X}_i$ is partitioned into M regions, that we denote $R_1, \ldots R_M$. We introduce the class F of constant piecewise functions on the elements of the partition :

$$F = \{f, f(\boldsymbol{x}) = \sum_{m=1}^{M} c_m \mathbf{1}_{\boldsymbol{x} \in R_m}\}.$$

The least square estimator of the regression function f on the class F minimizes the criterion

$$\sum_{m=1}^{M} (Y_i - f(\boldsymbol{X}_i))^2,$$

among the functions $f \in F$. The solution is

$$\hat{f}(\boldsymbol{x}) = \sum_{m=1}^{M} \hat{c}_m \mathbf{1}_{\boldsymbol{x} \in R_m},$$

where $\hat{c}_m$ is the mean of observations Y_i such that $\boldsymbol{X}_i \in R_m$. In order to define the partition, CART proceeds as follows : given a separation variable $X^{(j)}$ and a point of separation s, one considers the half spaces

$$R_1(j,s) = \{\boldsymbol{X} = (X^{(1)}, \ldots, X^{(d)}) / X^{(j)} \leq s\} \text{ and } R_2(j,s) = \{\boldsymbol{X} / X^{(j)} > s\}.$$

The separation variable $X^{(j)}$ and the separation point s are chosen in order to solve the minimisation problem

$$\min_{j,s} [\sum_{i, \boldsymbol{X}_i \in R_1(j,s)} (Y_i - \hat{c}_1)^2 + \sum_{i, \boldsymbol{X}_i \in R_2(j,s)} (Y_i - \hat{c}_2)^2].$$

Given j and s, we partition the data in the two corresponding regions, and then we proceed to a new separation on each of the two subregions, and so on, on each obtained subregion. The size of the tree is a parameter to adjust, that

is related to the complexity of the tree. A tree with a large number of leaves will lead to overfitting (large variance) and a tree with small size will lead to under-fitting (large bias). It is therefore necessary to find the optimal size of the tree with an adaptive procedure, driven from the data. The strategy is to build a large tree and to prune the tree by minimizing a penalized criterion.

We call T a sub-tree of T_0 if T can be obtained by pruning T_0, that is by reducing the number of nodes of T_0. We denote $|T|$ the number of terminal nodes of T and $R_m, m = 1, \dots |T|$, the partition corresponding to the terminal nodes. We denote by N_m the number of observations for which $\boldsymbol{X}_i \in R_m$. We have

$$\hat{c}_m = \frac{1}{N_m} \sum_{i, \boldsymbol{X}_i \in R_m} Y_i,$$

and we introduce the criterion to be minimized

$$C_\lambda(T) = \sum_{m=1}^{|T|} \sum_{i, \boldsymbol{X}_i \in R_m} (Y_i - \hat{c}_m)^2 + \lambda |T|.$$

For all λ, we can prove that there exists a unique minimal tree T_λ minimizing the criterion $C_\lambda(T)$. To find the tree T_λ, we suppress, at each step the internal node of the tree T which reduces the less the criterion $\sum_m \sum_{i, \boldsymbol{X}_i \in R_m} (Y_i - \hat{c}_m)^2$. This gives a sequence of subtrees, that contain the tree T_λ.

The regularization parameter λ must also be calibrated to realize a good compromise between the bias and the variance of the final estimator. We generally use cross validation, described in the next section, to calibrate this parameter.

Appendix : Choice of a tuning parameter by cross-validation

In the case of kernel estimators, for estimators based on local polynomials, we have to choose the window h; for constant piecewise estimators (or piecewise polynomials), as well as for projection estimators, we have to choose the parameter D (number of pieces of the partition or dimension of the linear space onto which the projection is realized), for CART algorithm, we have to choose the parameter λ for the pruning procedure. In this chapter, we are going to describe the cross-validation method, which is often used by the statistical softwares to tune these parameters, which corresponds to select an estimator among a collection of estimators.

We denote by λ the parameter to be tuned. Let $\hat{f}_{n,\lambda}$ the estimator of the regression function f associated to the parameter λ. We consider the quadratic risk

$$R(\lambda) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} (\hat{f}_{n,\lambda}(\boldsymbol{X}_i) - f(\boldsymbol{X}_i))^2 \right).$$

Ideally, we want to choose λ to minimize $R(\lambda)$, but this quantity depends on the unknown function f.

The first idea is to estimate $R(\lambda)$ by the **training error** :

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_{n,\lambda}(\boldsymbol{X}_i))^2,$$

but this quantity is optimistic, it underestimates $R(\lambda)$ and leads to overfitting. This is due to the fact that we use the same data to build the estimator $\hat{f}_{n,\lambda}$ (that is well adjusted to the data) and to estimate the risk for this estimator. In order to get a better estimation of the risk, we have to build the estimator of the risk with the observations that were not use to compute the estimator $\hat{f}_{n,\lambda}$. Ideally, if we have enough observations, we can separate them in a training sample and a test sample. This is generally not the case, and we would like to use all the data to build the estimator. In this case, we use cross-validation. We split the learning sample into V blocks, denoted by $B_1, \dots B_V$, with quite similar sizes. For v from 1 to V, we denote $\hat{f}_{n,\lambda}^{(-v)}$ the estimator obtained by removing from the learning sample the data from the block B_v.

DEFINITION 10. —

We define the V-fold cross-validation score by

$$CV = \hat{R}(\lambda) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{f}_{n,\lambda}^{(-v(i))}(\boldsymbol{X}_i))^2,$$

where $\hat{f}_{n,\lambda}^{(-v(i))}$ is the estimator of f obtained by removing the observations of the block that contain the observation $(\boldsymbol{X}_i, Y_i)$.

The principle of cross-validation is to choose a value $\hat{\lambda}$ of λ which minimizes the quantity $\hat{R}(\lambda)$. A particular case corresponds to the leave-one-out cross-validation, obtained when we consider n blocks, each of them containing a single observation.

DEFINITION 11. — *The leave-one-out cross-validation score is defined by*

$$CV = \hat{R}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{n,\lambda}^{(-i)}(\boldsymbol{X}_i))^2,$$

where $\hat{f}_{n,\lambda}^{(-i)}$ is the estimator of f obtained by removing the observation $(\boldsymbol{X}_i, Y_i)$.

The idea of the leave-one-out cross validation comes from the following computation :

$$\begin{aligned}\mathbb{E}((Y_i - \hat{f}_{n,\lambda}^{(-i)}(\boldsymbol{X}_i))^2) &= \mathbb{E}((Y_i - f(\boldsymbol{X}_i) + f(\boldsymbol{X}_i) - \hat{f}_{n,\lambda}^{(-i)}(\boldsymbol{X}_i))^2)\\ &= \sigma^2 + \mathbb{E}((f(\boldsymbol{X}_i) - \hat{f}_{n,\lambda}^{(-i)}(\boldsymbol{X}_i))^2)\\ &\simeq \sigma^2 + \mathbb{E}((f(\boldsymbol{X}_i) - \hat{f}_{n,\lambda}(\boldsymbol{X}_i))^2).\end{aligned}$$

We then obtain $\mathbb{E}(\hat{R}(\lambda)) \simeq \sigma^2 + R(\lambda)$.

The computation of $\hat{R}(\lambda)$ may be very long but it is sometimes note necessary to compute n estimators of the regression function. For most regression estimation methods considered in this chapter, the estimator corresponds to a local mean algorithm, namely it can be written as follows

$$\hat{f}_{n,\lambda}(\boldsymbol{x}) = \sum_{j=1}^{n} Y_j l_j(\boldsymbol{x}),$$

with $\sum_{j=1}^{n} l_j(\boldsymbol{x}) = 1$. One can prove that

$$\hat{f}_{n,\lambda}^{(-i)}(x) = \sum_{j=1}^{n} Y_j l_j^{(-i)}(\boldsymbol{x}),$$

with

$$\begin{aligned}l_j^{(-i)}(\boldsymbol{x}) &= 0 \text{ if } j = i\\ &= \frac{l_j(\boldsymbol{x})}{\sum_{k\neq i} l_k(\boldsymbol{x})} \text{ if } j \neq i.\end{aligned}$$

THEOREM 12. — *Under the above assumptions, the leave-one-out cross validation score equals*

$$CV = \hat{R}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{f}_{n,\lambda}(\boldsymbol{X}_i)}{1 - l_i(\boldsymbol{X}_i)}\right)^2.$$

We also find in the softwares a slightly different definition :

DEFINITION 13. — *We called the Generalized Cross Validation score the quantity*

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{f}_{n,\lambda}(\boldsymbol{X}_i)}{1 - \nu/n}\right)^2,$$

where $\nu/n = \sum_{i=1}^{n} l_i(\boldsymbol{X}_i)/n$.

In this definition, $l_i(\boldsymbol{X}_i)$ is replaced by the mean of the quantities $l_i(\boldsymbol{X}_i)$. In practice, both methods generally give quite similar results. Using the approximation $(1-x)^{-2} \approx 1 + 2x$ for x close to 0, we get

$$GCV(\lambda) \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{n,\lambda}(\boldsymbol{X}_i))^2 + \frac{2\nu\hat{\sigma}^2}{n},$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{n,\lambda}(\boldsymbol{X}_i))^2$. This corresponds to the Mallow's C_p criterion.

References

[1] Christophe Giraud. *Introduction to high-dimensional statistics*, volume 139 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015.

[2] T. Hastie, R. Tibshirani, and J Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer, 2009. Second edition.

[3] P.. Massart. *Concentration inequalities and Model selection*. Springer, 2003.

[4] A. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009.